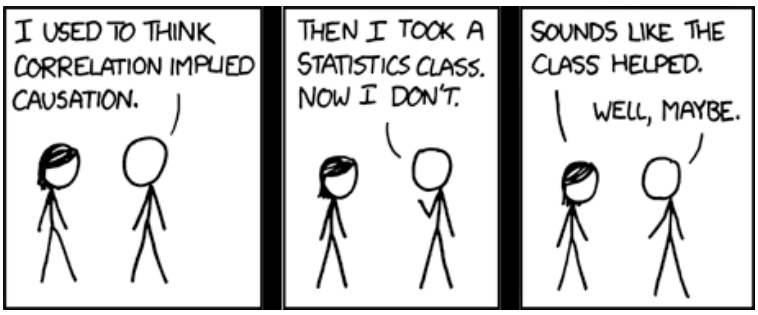
# Correlation, causation, and sample size




---

**43 marks**

In the May/June 2013 issue of the publication *Foreign Affairs*, Kenneth Cukier (Data Editor of *The Economist*) and Viktor Mayer-Schoenberger (Professor of "Internet Governance and Regulation" at the *Oxford Internet Institute*) wrote a paper called "The Rise of Big Data: How It's Changing the Way We Think About the World".

There, they write that modern times require "three profound changes in how we approach data.

> "The first is to collect and use a lot of data rather than settle for small amounts or samples, as statisticians have done for well over a century. The second is to shed our preference for highly curated and pristine data and instead accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. Third, in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations. With big data, instead of trying to understand precisely why an engine breaks down or why a drug's side effect disappears, researchers can instead collect and analyze massive quantities of information about such events and everything that is associated with them, looking for patterns that might help predict future occurrences. Big data helps answer what, not why, and often that's good enough."

In has become a popular sentiment that "the more data the better," even if it is of "variable quality" and whatever its "messiness" might be. The what, if not the why, of a causal relation might yet be determined from such data.

To investigate these ideas, you will be exploring some of the data that was measured in the physical laboratory. (Background on the lab and its experimental protocols are described in the document `labBackground`.) Three variables $x$, $y$, and $z$, are linearly related (lie on a plane) and the value of any two determine (or cause) the value of the third. That is,

$$y = \alpha + \beta x + \gamma z$$

Changing only one of the variables, $x \to x + \Delta x$, while holding a second, $z$, fixed at some value, will *cause* the third to change, $y \to y + \Delta y$ in response. We understand that changes in $x$ *cause* changes in $y$. Moreover, in this instance, that change is $\Delta y = \beta \times \Delta x$ and $x$ can cause changes in $y$ if and only if $\beta \neq 0$. If non-zero, the sign and magnitude of $\beta$ determine the nature of the causal relationship.

Typically not all of the variables in a causal relationship are recorded. In the physical laboratory, only values of $x$ and $y$ were collected; no values of $z$ (though these too were changing). What separated the `type`s was *how the data were collected*.

Of interest, is to determine what we can learn from the data, if anything, about the value of $\beta$ and hence about the causal relationship between $x$ and $y$.

# 1 Data

Set up the following:

```
## Set this up for your own directory
imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
```

and then read in the data as:

```
labData <- read.csv(file = path_concat(dataDirectory, "labData.csv"))
```

Interest here lies **only** in that subset of the data where:

- `type` is either `"observational"` **or** `"randomized"`, **and**
- `rep` is 1

The result should be a `data.frame` having five variables and 216 rows (108 for each `type`). The value of `rep` should be 1 for all rows and the value of `type` should only be one of the two mentioned above.

a. *(2 marks)* Construct the `data.frame` described above. Assign it to the variable `results`. Show your code.

# 2 Analysis

The dataset `results` from part (a) contains 216 $(x, y)$ pairs, half of which are of `type "observational"` and half of which are of `type "randomized`". To address the problem of whether changes in $x$ cause changes in $y$ you will pursue fitting a straight line model to the data of the form

$$y = \beta_0 + \beta_1 x + r$$

where $r$ is the residual representing the fact that $y$ and $x$ need not lie exactly on a line. All fitting will be done by least-squares using the `lm()` function from R .

Of interest will be the hypothesis $H_0 : \beta = 0$. As a surrogate, we will use the model to test $H_0 : \beta_1 = 0$ and to provide estimates for $\beta_1$ in place of estimates of $\beta$.

**Note** that parts (b), (c), and (d) are identical, the difference lies in which data are used.

b. Following Cukier's and Mayer-Schoenberger's advice, here you will use **all** of the data in `results`.

   i. *(2 marks)* Fit a straight line model of $y$ to $x$ and print the `summary()` of the fitted model. Show your code.

   ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis $H_0 : \beta_1 = 0$? Justify your answer.

   iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer.

   iv. *(2 marks)* Repeat part (iii) above but now ise the absolute value of the sample correlation as the discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer(s).

   v. *(2 marks)* What do the above empirical tests, based on **all** the data here, suggest one might conclude about the causal relation between $x$ and $y$? If you have found a causal effect, in what direction would $y$ be expected to change if a value of $x$ were increased? Justify your answers.

c. Now restrict the analysis to only that subset of `results` which has `type "observational"`. This analysis will be based on only 108 $(x, y)$ pairs.

   i. *(2 marks)* Fit a straight line model of $y$ to $x$ and print the `summary()` of the fitted model. Show your code.

   ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis $H_0 : \beta_1 = 0$? Justify your answer.

   iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer.

   iv. *(2 marks)* Repeat part (iii) above but now ise the absolute value of the sample correlation as the discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer.

   v. *(2 marks)* What do the above empirical tests, based on **only** the `"observational"` data here, suggest one might conclude about the causal relation between $x$ and $y$? If you have found a causal effect, in what direction would $y$ be expected to change if a value of $x$ were increased? Justify your answers.

d. Now restrict the analysis to only that subset of `results` which has `type "randomized"`. This analysis will be based on only 108 $(x, y)$ pairs.

   i. *(2 marks)* Fit a straight line model of $y$ to $x$ and print the `summary()` of the fitted model. Show your code.

   ii. *(1 mark)* What do you conclude from this summary about the evidence against the hypothesis $H_0 : \beta_1 = 0$? Justify your answer.

   iii. *(3 marks)* Using the `numericalTest()` function, perform a test of the hypothesis of independence (that is, $H_0 : X \perp\!\!\!\perp Y$) based on the absolute value of the slope estimate as a discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer.

   iv. *(2 marks)* Repeat part (iii) above but now ise the absolute value of the sample correlation as the discrepancy measure. Show your code.

   What do you conclude about the hypothesis? Justify your answer.

   v. *(2 marks)* What do the above empirical tests, based on **only** the `"randomized"` data here, suggest one might conclude about the causal relation between $x$ and $y$? If you have found a causal effect, in what direction would $y$ be expected to change if a value of $x$ were increased? Justify your answers.

# 3 Conclusions

You have now conducted three separate analyses of the data. Based on the results in these analyses, address the following.

e. *(4 marks)* What is meant by a "lurking" variable and how might it have shown up in this data?

f. *(3 marks)* What conclusion do you draw about the causal relation, if any, between changes in $x$ and changes in $y$? Justify your answer.

g. *(2 marks)* In 2008, then editor of *Wired* magazine, Chris Anderson wrote in an article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete":

*There is now a better way. Petabytes allow us to say: "Correlation is enough."*

Drawing on your analyses above, give an illustration as to why Chris Anderson might be mistaken, in spite of the volume of data observed.

h. *(2 marks)* Why might Cukier's and Mayer-Schoenberger's advice (quoted above) actually be dangerous?